## Sept. 2017

## Gathering Data for your project P2

(Revised from Jim Noble, In-Thinking Site)

In order to carry out your project you will need to have some information, measurements or data about a topic of interest to you. Once your topic is chosen, and information/data is gathered, you then need to be very specific about how it relates to your topic. Before you make decisions and answer questions about your information/data, you should consider the following list of points:

* **Your data must not be one dimensional** (ie: temperature change over the last 50 years, lots of numbers but they are all temperatures. A better choice would be to add more dimension by comparing different regions, comparing environmental changes connected to temperature change, etc…).
* **Your data must be largely numerical** - categorical fields are important as well but not enough on their own. (Hair color and gender are both only categorical. At least include some other numerical data to complement it).
* **Your data must be sufficient in quantity** – ie: a survey might be best with a minimum of 100 to 200 responses.
* **If you find data on the internet, it must be easily transferred into a spreadsheet for huge time saving and flexibility benefits.** This may be possible by copying and pasting into a spreadsheet program like Excel or Google Sheets.
* **If you plan a survey, you must spend time working soundly on your questionnaire**;
* You can not go back and add/change questions once you have done it.
* Where possible, surveys should involve some kind of measurement.
* Information collected experimentally or with surveys needs to be collected under defined and consistent conditions and you are required to describe these conditions.
* **Where possible, it is good if you can use your information to generate new information** (for example, collect height and weight to generate body mass index).

**How, what and where?**

Information collection is central to this project. If the information is not sufficient in ***quality*** or ***quantity*** then the rest of the project will suffer accordingly. Information for projects is often of a statistical nature and so there is an emphasis on how you will make sure it is ***sufficient quality and quantity***. If you choose a non-statistical project the points are still relevant as you will have to make careful decisions about which calculations to make or which information to use for modeling. The following is a set of guidelines to help make sure that quality and quantity give sufficient potential for investigation whatever the nature of the project.

### Sufficient Quality & Quantity

The definition of the these two words will vary depending on the project, so it is difficult to give a hard and fast guideline on what is sufficient. It is important to consider the nature of the information. It is very difficult (and not appropriate) to perform mathematical analysis without enough data and equally difficult to search for interesting conclusions without having multi-variable and rich data. Strive for a balance between these two. For example, make sure that there is enough categorical data to categorize the data set in interesting ways so that the associated numerical data can be compared accordingly. There is a requirement that information not be too *'one dimensional'* and so variety in both types of information is recommended.

#### Comparisons

*More often than not statistics have greater meaning when compared to the same statistic for a different data set. For example it is always interesting to compare the average cancer rates in a certain region to the world average or those results from previous years in the same region.*

*You should try to look for ways to categorize your information so that such comparisons can be made. This implies thought before collection! This also helps you make interesting interpretations.*

**A. Primary Data (Data you personally collect)**

Data from questionnaires, results of some experiments, measurements, or calculations, etc…

The major advantage of collecting primary data for projects is that it gives you total control over what information you will have. Primary data is data that you personally collect (or measure, etc…). You do not have to worry about whether you will find what you want to analyze in a pre existing data set. Formulate your ideas and then decide what information you need to collect in order to prove your point. Sometimes the data will lead to surprising results that lead you to different conclusions than you expected which leads to rich analysis!

The down side is that collecting primary data can be time consuming. First, deciding which information to collect and then collecting it. Even after significant time input, you may still not have the quantity or quality needed. Added to that there is no obvious extra credit given in the grading criteria for doing it this way. It can be frustrating with the associated limits of a small data set because there is only so much you can do with it.

A key point to remember is that when collecting primary data it is often very difficult or impossible to collect an extra field heading to a database *retrospectively*! For example, having asked 200 people 20 questions it would be very difficult and time consuming to ask them all a new question you thought of afterwards. With this in mind it is very important to make sure the decisions on what is to be collected are clearly considered in advance!

#### Experiments and Measurements. Sample ideas:

* **human proportions**: when one measures different dimensions of the human body on a sample of students for example, then compares them.
* **human testing**, where different students are measured and timed doing different activities from running and jumping, to weight lifting or even the time it takes to complete different puzzles.

There are many more experimental ideas that can lead to statistical or modeling projects. The key issues with this type of information collection are:

* **Sample size** - How many measurements are needed? There is a nominal figure of 50 as a minimum but that can be limiting, depending on the nature of the information and often 200 is a much more interesting and flexible figure, while 1000 might generate some very interesting possibilities. In general the answer is going to be 'as many as possible' before the time spent collecting information begins to encroach on time required to complete the rest of the project.
* **Planning** - there is a time implication here. How long does it take to time 200 people doing 5 different activities and collect the information about them that allows them to be categorized? If you plan to measure the temperature at 5 times every day in the same place for a period of 50 days then you are going to need a head start! Time implications are very easily underestimated and this is often the biggest reason for last minute pressure build up and stress to finish.
* **Equipment** - again, this is often underestimated. It is easy to say, I'll get a this and I'll get a that and when you find you are not able to get exactly what you need, you compromise your information collection and your project.
* **Consistency** - Experimental conditions should, of course, be the same for the collection of all the information. In practice this is not always easy but you should attempt to do so as much as possible and note when you have not been able to keep the experimental conditions the same. Use this as part of your discussion in writing up the project. Such conditions can be anything from 'time of day' to 'degree of accuracy'.

#### Questionnaires

Questionnaires can lead to some of the best and worst projects. They should always be considered an option, but only if you have a plan that shows how the Mathematical analysis of the results will be relevant to the goals of the task. “*Questionnaires can be appealing and*with*the aforementioned relevance they can lead some absolutely brilliant projects*”.

The same issues apply regarding **sample** and **planning** for questionnaires as apply for experiments and measuring. In addition, students should consider the following:

* **Technology** - data collection tools such as 'Google forms' and 'Survey Monkey' take a huge amount of the legwork out of the collection and organization of responses. Students still have to make the key strategic decisions relating to the questions and how they are answered. The process of making the questionnaires helps with these decisions too because you have to decide how the question will be answered in advance. A student can complete a questionnaire and then man a station with one or more computers during breaks and lunchtimes while students fill in and submit the answers. E-mail allows the questionnaire to reach much further afield and helps students question larger samples of people. Once submitted the responses are automatically compiled in a spreadsheet.
* **Numerical responses** - as mentioned above, students need to keep a balance of numerical and categorical data to make sure they have enough numbers to perform analysis. Students can do this by asking questions with numerical answers. One method is to ask questions that ask for 'numerical ratings'.
* **Testing the Questionnaire** - Students must spend time testing their questionnaire and will often find that they have 3 or 4 drafts before they have the questionnaire they want. They should show these drafts to others and look for improvements and ask someone to take the questionnaire and then comment on what they found difficult to answer or unclear. Students may also try a small sample of 10 students to get a preliminary idea of how it will go. They can then adjust the questions accordingly before using it on a larger sample of people.

**B. Secondary Data**

Using secondary data can have the advantage of saving huge amounts of time and allowing for considerably larger data sets from broader sources that can lead to more interesting investigations, conclusions and interpretations. The Internet is an incredible source of such information and potentially a major tool. Many projects would simply not be possible by using primary data alone. The use of secondary data opens up many more possibilities, which may help you to find an idea you are really interested in.

#### Sources

Sources have to, of course, be credited and referenced. Equally, you should spend some time questioning the reliability of the chosen source by examining how and when the data was collected and what the associated limitations might have been. This can make for interesting discussion in your written reports and should not necessarily stop you using the source. One of the biggest problems occurs when one uses multiple sources to build a single data set. The conditions under which the data was originally collected become increasingly relevant if the two data sets are to be considered together. Again, this should be looked at and written about, without stopping you from using the multiple sources. The other associated snag with multiple sources is that you can only really using the field headings that the two sources have in common and so this should be taken into account when considering the different sources. Ultimately the choice of using different sources can effect how much time is actually saved by using secondary data!

“A further warning on secondary data is that it is often difficult to find 'Raw Data' that is published. You are much more likely to find somebody's summary statistics that hint at the existence of the raw data you are looking for. This can make the search even more difficult and students should be made aware of this distinction.”

#### Formatting

Formatting seems a trivial issue but can actually have a major bearing on how useable the data is. If you find a database with over 400 entries on the Internet for example, but you can't easily import it into a spreadsheet then you are left with very limited possibilities or a desperately painstaking process of copying data over in small steps.

### C. Modeling

There are numerous possibilities for projects on modeling, where information collected is modeled by a function. Examples include:

* Athletic stats - exponential growth and decay
* Paths of projectiles - quadratics
* Temperature fluctuations - sine waves

These can make for some very interesting projects, but you should, again, focus on making sure you have a systematic approach to collecting the information and choosing your model(s). There should be some discussion of the limits of your model in making further predictions. Based on the three ideas above, here are some considerations:

* *When looking at athletic events, how do you choose the events to look at? Would the 100m, 200m, 400m, 800m, 1500m be more interesting to look at as a sequence than a random collection of events?*
* *Consider the quadratic path of a rugby ball being kicked between the posts from different parts of the pitch. How are the positions chosen? What are the common conditions? How can one introduce a systematic sequence into the way they collect such information?*
* *When comparing the temperature fluctuations of different cities, how does one choose the cities? What is the logical progression from one comparison to the next?*

With such considerations, these projects can reach some lofty heights. It is important that you get as much potential from the information collection as possible even if that potential is not eventually realized.

### Combination of Primary and Secondary Data

Combining some or all of the above approaches to information collection could go either way. You could end up with an unrelated hodgepodge of numbers or a well rounded, coherent set of relevant comparisons. It is not being suggested here that you combine these methods, but only that you could. The following two features can be particularly desirable in projects:

* *Collect primary data and find a suitable secondary data source with which to compare your findings.*
* *Use either the primary data or secondary data to make calculations that generate new field headings.*